Claims

We Claim:

1. In a system having a plurality of computers each having data sets stored thereon, a method of assigning a computer to service a request for a data set, said method comprising the steps of:

(a) providing a neural network having at least an input layer having J input nodes and an output layer having K output nodes, each of said output nodes associated with one of said computers, and associated weights w(j,k) between each said input node and each said output node;

(b) receiving a request for particular data set I;

(c) imputing to said input layer an input vector having an entry R(I) at input node I, said entry R(I) being dependent upon the number of requests for the requested data over a predetermined period of time; and

(d) selecting a computer assignment associated with a selected one of said output nodes to service said data request, where said selected output node is associated with a specific weight, said specific weight selected to minimize a predetermined metric measuring the distance between said vector entry R(I) and the weights(I,k) associated with said input node I and said output nodes.

2. The method of claim 1 where said method further includes the step of updating said specific weight.

3. The method of claim 2 where said step of updating said specific weight includes modifying said specific weight with a factor dependent said metric distance between said vector entry R(I) and said specific weight.

4. The method of claim 3 where said step of updating said specific weight further includes modifying said specific weight with a means to balance the load across a subset of said output nodes.

5. The method of claim 4 where said means to balance the load across a subset of said output nodes is dependent upon the number of data requests serviced by said subset of said output nodes over said predetermined period of time divided by the number of output nodes in said subset of said output nodes.

6. The method of claim 2 wherein R(I) is proportional to the ratio of (the number of previous requests for the requested data set) and (the number of previous requests for a subset of all data sets), over said predetermined period of time.

7. The method of claim 2 wherein each output node is associated with a neighborhood of other output nodes, and said step of updating said specific weight includes updating each weight in said neighborhood of said output node associated with said specific weight.

8. The method of claim 2 where said update is according to the formula $W(I,j)=W(I,j) +alpha((R(I)-w(I,j)) + beta(\sum W(i,k) -gama*W(I,j))$, where alpha, beta and gama are pre-determined constants.

9. The method of step 1 where said input vector's components, other than said component R(I) associated with said input node I, are of value zero.

10. In a web farm of servers, a method of selecting a server to service a user request for a data set comprising the steps of:

(a) providing a neural network having at least an input layer having J input nodes and an output layer having K output nodes, each of said output nodes associated with one of said servers, and associated weights w(j,k) between each said input node and each said output node;

(b) receiving a request for particular data set I;

(c) imputing to said input layer an input vector having an entry R(I) at input node I, said entry R(I) being dependent upon the number of requests for the requested data over a predetermined period of time,

(d) selecting a server assignment associated with of one of said output nodes to service said data request, where said output node is associated with a specific weight, said specific weight selected to minimize a predetermined metric measuring the distance between said vector entry R(I) and the weights(I,k) associated with said input node I and said output nodes.

11. A method implemented in a web farm according to claim 11, where said method is implemented on at least one server in said web farm.

12. A method implemented in a web farm according to claim 11 where said method is implemented on at least one router in said web farm.

13. The method according to claim 1 further comprising the step of transmitting said request to said server associated with said server assignment.

14. A computer readable storage medium containing computer executable code for performing a method of assigning a computer from a set of computers to service a request for a data set, said method comprising the steps of:

(a) associating for each data set I a series of weights w(I,j), where j=1,number of computers in the set of computers, associating with each individual weight w(I,j) one of said computers from said set of computers;

(b) receiving a request for particular data set I;

(c) associating with said requested data set a value R(I) being dependent upon the number of requests for the requested data set over a predetermined period of time,

(d) selecting a computer assignment associated with a specific one of said series of weights w(I,j) to service said data request, where said specific weight is selected to minimize a predetermined metric measuring the distance between said value R(I) and the weights(I,k) associated with said particular data set I.

15. A computer readable storage medium containing computer executable code for performing a method of assigning a computer for a set of computers to service a request for a data set, said method comprising the steps of:

(a) providing a neural network having at least an input layer having J input nodes and an output layer having K output nodes, each of said output nodes associated with one of said computers, and associated weights w(j,k) between each said input node and each said output node;

(b) receiving a request for particular data set I;

(c) imputing to said input layer an input vector having an entry R(I) at input node I, said entry R(I) being dependent upon the number of requests for the requested data over a predetermined period of time,

(d) selecting a computer assignment associated with of one of said output nodes to revise said data request, where said output node is associated with a specific weight, said specific weight selected to minimize a predetermined metric measuring the distance between said vector entry R(I) and the weights(I,k) associated with said input node I and said output nodes.